Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning

Year: 2023 | *Citations: 393* | *Authors: Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng*

Abstract

The popularity of LLaMA (Touvron et al., 2023a;b) and other recently emerged moderate-sized large language models (LLMs) highlights the potential of building smaller yet powerful LLMs. Regardless, the cost of training such models from scratch on trillions of tokens remains high. In this work, we study structured pruning as an effective means to develop smaller LLMs from pre-trained, larger models. Our approach employs two key techniques: (1) targeted structured pruning, which prunes a larger model to a specified target shape by removing layers, heads, and intermediate and hidden dimensions in an end-to-end manner, and (2) dynamic batch loading, which dynamically updates the composition of sampled data in each training batch based on varying losses across different domains. We demonstrate the efficacy of our approach by presenting the Sheared-LLaMA series, pruning the LLaMA2-7B model down to 1.3B and 2.7B parameters. Sheared-LLaMA models outperform state-of-the-art open-source models of equivalent sizes, such as Pythia, INCITE, OpenLLaMA and the concurrent TinyLlama models, on a wide range of downstream and instruction tuning evaluations, while requiring only 3% of compute compared to training such models from scratch. This work provides compelling evidence that leveraging existing LLMs with structured pruning is a far more cost-effective approach for building competitive small-scale LLMs